CLAIMS

1. A proxy partition cache architecture for vending objects to requesting clients comprising:

a plurality of proxy cache servers for caching and vending objects from a storage location to a requesting client through an interconnected load-balancing mechanism adapted to selectively provide client requests to each of the plurality of proxy cache servers based upon load-balancing considerations;

a network segment interconnecting each of the proxy cache servers so as to enable data to pass between the proxy cache servers; and

a mechanism in each server of the plurality of proxy cache servers adapted to (a) receive a request for an object from the load-balancing mechanism, (b) identify a discrete server of the plurality of proxy cache servers that is designated to contain the object, based upon an address of the object, (c) referring the request to the discrete server and (d) return the object from the discrete server to the server receiving the request from the load-balancing mechanism for vending to the client.

2. The proxy partition cache architecture as set forth in claim 1 wherein the mechanism includes means for the request to the discrete server to be referred over the network segment unless the discrete server and the receiving server are identical, whereby the request is optimally processed on the receiving server.

3. The proxy partition cache architecture as set forth in claim 1 wherein the mechanism includes means for performing a hash function on the address of the object.

4. The proxy partition architecture as set forth in claim 3 wherein the mechanism includes a means for performing a modulo function on a hash of the address with respect to a number of proxy cache servers in the plurality of proxy cache servers.

5. The proxy partition architecture as set forth in claim 1 wherein the load-balancing mechanism comprises a network switch interconnected to the proxy cache servers.

6. The proxy partition architecture as set forth in claim 1 wherein the mechanism includes means for delivering the request to the discrete server and for tunneling the object through the server receiving the request from the discrete server to the load-balancing mechanism.

7. The proxy partition architecture as set forth in claim 1 wherein the mechanism includes means for enabling a redirection of the client request to the discrete server from the server receiving the request from the load-balancing mechanism.

8. The proxy partition architecture as set forth in claim 7 wherein the mechanism includes means for notifying the discrete server to receive the referred request so that a client redirection is expected, and allowing the discrete server to reject the request when the request is not recognized as a referred request by the mechanism.

9. The proxy partition architecture as set forth in claim 1 wherein the mechanism includes means for allowing vending of a file from a server other than the one selected by the modulo mechanism based upon a directory entry that indicates a current server location and data freshness of the file.

10. A method for address-partitioning a proxy cache defined by a group of proxy cache servers for vending objects to requesting clients comprising:

caching and vending objects from a plurality of proxy cache servers to a requesting client through an interconnected load-balancing mechanism, including selectively providing client requests to each of the plurality of proxy cache servers based upon load-balancing considerations;

interconnecting each of the proxy cache servers through a network segment so as to enable data to pass between the proxy cache servers; and

in each server of the plurality of proxy cache servers, (a) receiving a request for an object from the load-balancing mechanism, (b) identifying a discrete server of the plurality of proxy cache servers that is designated to contain the object, based upon an address of the object, (c) referring the request to the discrete server over the network segment and (d) returning the object from the discrete server to the server receiving the request from the load-balancing mechanism for vending to the client.

11. The method as set forth in claim 10 further comprising referring the request to the discrete server over the network segment unless the discrete server and the receiving server are identical, whereby the request is optimally processed on the receiving server.

12. The method as set forth in claim 10 wherein the step of identifying includes performing a hash function on the address of the object.

13. The method as set forth in claim 12 wherein the step of identifying includes performing a modulo function on a hash of the address with respect to a number of proxy cache servers in the plurality of proxy cache servers.

14. The method as set forth in claim 10 wherein the load-balancing mechanism comprises a Layer 4 (L4) switch.

15. The method as set forth in claim 10 wherein the steps of referring and returning each include delivering the request to the discrete server and tunneling the object through the server receiving the request from discrete server to the load-balancing mechanism.

16. The method as set forth in claim 10 further comprising allowing vending of a file from a server other than the one selected by the modulo mechanism based upon of a directory entry that indicates a current server location and data freshness of the file.

17. The method as set forth in claim 10 wherein the steps of referring and returning include redirecting of the client request to the discrete server from the server receiving the request from the load-balancing mechanism.

18. The method as set forth in claim 17 further comprising notifying the discrete server to receive the referred request so that a client redirection is expected, and allowing the discrete server to reject the request as an unallowed external request when the request is not recognized as allowable.

19. A computer-readable medium including program instructions for address-partitioning a proxy cache defined by a group of proxy cache servers for vending objects to requesting clients, the computer-readable medium including instructions for performing the steps of:

caching and vending objects from a plurality of proxy cache servers to a requesting client through an interconnected load-balancing mechanism, including selectively providing client requests to each of the plurality of proxy cache servers based upon load-balancing considerations;

interconnecting each of the proxy cache servers through a network segment so as to enable data to pass between the proxy cache servers; and

in each server of the plurality of proxy cache servers, (a) receiving a request for an object from the load-balancing mechanism, (b) identifying a discrete server of the plurality of proxy cache servers that is designated to contain the object, based upon an address of the object, (c) referring the request to the discrete server over the network seg-

ment and (d) returning the object from the discrete server to the server receiving the request from the load-balancing mechanism for vending to the client